# Nutrition-Aware Food Image Classification Using Vision Transformers on Food-101

Yuze Jin (54587468)
STATS 507 – Data Science Analytics using Python
University of Michigan, Ann Arbor, MI, USA

***Abstract*—Accurate dietary tracking plays an essential role in promoting long-term health, yet many nutrition applications still require users to manually enter food information. This process is time-consuming and error-prone, often resulting in unreliable nutritional logs. Image-based food recognition offers a promising alternative due to the ubiquity of smartphone cameras. However, food classification remains challenging because of extreme visual variability, ambiguous category boundaries, and cluttered backgrounds.**

**In this project, I develop an end-to-end nutrition-aware food recognition system using the Food-101 dataset and a pretrained Vision Transformer (ViT-Base/16). I implement a full training pipeline using Hugging Face `datasets` and `transformers`, perform fine-tuning with data augmentation, evaluate model performance, and analyze misclassification patterns. The final model achieves 87.05% Top-1 and 97.50% Top-5 accuracy on the test split ($n = 25{,}250$). I further incorporate a lightweight nutritional lookup table based on USDA FoodData Central, demonstrating how classification outputs can be mapped to approximate macronutrient estimates. Confusion-pair visualization shows that most errors arise from subtle inter-class similarities or multi-item images. These findings illustrate both the strengths and limitations of transformer-based models for nutrition-facing applications.**

***Index Terms*—Food image classification, Vision Transformer, Food-101, Hugging Face, nutrition estimation.**

## I. Introduction

Maintaining an accurate food diary is one of the most effective strategies for long-term health management. However, manual logging of food names, ingredients, and serving sizes is burdensome and often inconsistent. Because taking a photo requires minimal user effort, image-based food recognition has become an appealing alternative.

Yet, food classification presents unique challenges. Dishes exhibit substantial intra-class variation depending on preparation style, lighting, angle, and plating. Conversely, different dishes may appear visually similar. Many images also contain multiple food items, yet datasets such as Food-101 assign only a single label. These difficulties require models capable of robust generalization.

Deep learning models, including convolutional neural networks (CNNs) [2], have historically performed well in food recognition tasks. More recently, the Vision Transformer (ViT) [3] has emerged as a powerful architecture that leverages global self-attention, making it suitable for food imagery where discriminative cues may be spatially distributed.

Beyond recognition, nutrition-oriented applications require connecting predicted categories to nutrient information. Although precise nutrient estimation from images is extremely challenging, mapping predicted classes to representative values from standardized databases such as USDA FoodData Central [4] allows me to demonstrate an interpretable proof of concept.

This project focuses on three objectives:

- Implement and fine-tune a ViT-Base/16 model on Food-101 using a reproducible Hugging Face workflow.
- Evaluate accuracy, analyze error patterns, and visualize frequent confusion pairs.
- Demonstrate a nutrition-aware pipeline by linking predictions to USDA nutrient profiles.

## II. Method

### A. Problem Formulation

Let $x \in \mathbb{R}^{224\times224\times3}$ denote an RGB food image and $y \in \{1, \ldots, 101\}$ its class label. The model $f_\theta$ outputs probabilities $\hat{p}$ via a softmax layer. I train the model by minimizing cross-entropy:

$$\mathcal{L}(\theta) = -\log \hat{p}_y. \tag{1}$$

At inference, the predicted class is $\hat{y} = \arg\max_i \hat{p}_i$, which is then mapped to nutrient values.

### B. Dataset and Splits

Food-101 [1] provides 101 classes with 1,000 images each. Using Hugging Face `datasets`, I:

- load the dataset and keep the official test split for final evaluation,
- create a 90/10 stratified validation split from the training set using seed 507.

### C. Preprocessing and Augmentation

All images are converted to RGB and resized to $224 \times 224$. I apply augmentation only to the training set, including:

- random resized cropping,
- horizontal flipping,
- mild color jitter,
- small-angle rotation.

Normalization follows ImageNet statistics.

### D. Vision Transformer Fine-Tuning

I fine-tune `google/vit-base-patch16-224-in21k`, which divides each image into $16 \times 16$ patches and processes them via multi-head self-attention. My training configuration is:

- learning rate = $5 \times 10^{-5}$,
- batch size = 16,
- weight decay = 0.05,
- 1 training epoch,
- mixed precision with `fp16`.

Although one epoch is short, ViT models transfer well to downstream tasks, allowing meaningful results even under compute constraints.

### E. Nutrition Mapping

To provide interpretable feedback, I construct a lookup table that links each Food-101 class to USDA nutrient data, including calories and macronutrients per 100g. While approximate, this module demonstrates how image recognition can support nutrition-oriented output.

## III. Results

### A. Accuracy and Evaluation

The final model achieves:

- **Top-1 accuracy: 87.05%**,
- **Top-5 accuracy: 97.50%**,
- test set size: 25,250 images.

These results show that ViT performs strongly on food imagery even with limited training time.

### B. Confusion Analysis

To understand failure modes, I analyze misclassified samples and identify the most frequent confusion pairs. Fig. 1 shows the top confusion patterns.

Common patterns include:

1) visually similar categories (e.g., steak vs. filet mignon),
2) emphasis on the most salient ingredient rather than the labeled dish,
3) difficulty with multi-item scenes.

### C. Nutrition Output Demonstration

For each prediction, the system retrieves a nutrient profile from the lookup table. Although it does not estimate serving size or recipe variation, this step connects classification outputs to actionable nutrition information.

## IV. Conclusion

I develop a complete workflow for nutrition-aware food classification using a Vision Transformer and the Food-101 dataset. The ViT-Base/16 model achieves strong accuracy, and confusion analysis reveals challenges unique to food imagery. The nutrition lookup module demonstrates how classification results can support health applications. Future work includes multi-label modeling, portion-size estimation, and ingredient-level understanding.

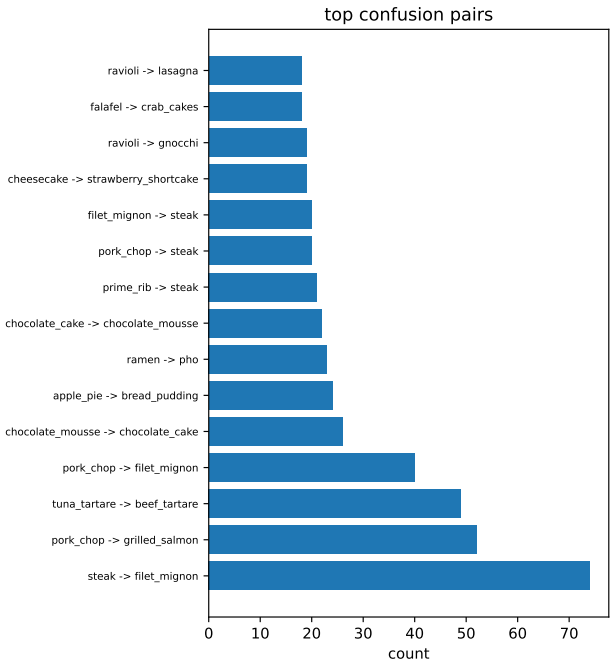


Fig. 1. Most frequent confusion pairs on the Food-101 test set.

## References


[1] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101: Mining discriminative components with random forests," *ECCV*, 2014.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, 2016.

[3] A. Dosovitskiy *et al.*, "An image is worth 16×16 words," *ICLR*, 2021.

[4] U.S. Department of Agriculture, "FoodData Central," https://fdc.nal.usda.gov.

[5] Hugging Face Documentation, https://huggingface.co/docs.